An L1 criterion for dictionary learning by subspace identification

Florent Jaillet, Rémi Gribonval, Mark D. Plumbley, Hadi Zayyani

AN L1 CRITERION FOR DICTIONARY LEARNING BY SUBSPACE IDENTIFICATION

Florent Jaillet [1], *Rémi Gribonval* [1], *Mark D. Plumbley* [2], *Hadi Zayyani* [3]

[1] Projet METISS, Centre de Recherche INRIA Rennes - Bretagne Atlantique,
IRISA, Campus de Beaulieu, F-35042 Rennes Cedex, France
[2] Queen Mary University of London, School of Electronic Engineering and Computer Science
Mile End Road, London E1 4NS, United Kingdom
[3] Sharif University of Technology, Department of Electrical Engineering, Tehran, Iran

ABSTRACT

We propose an ℓ^1 criterion for dictionary learning for sparse signal representation. Instead of directly searching for the dictionary vectors, our dictionary learning approach identifies vectors that are orthogonal to the subspaces in which the training data concentrate. We study conditions on the coefficients of training data that guarantee that ideal normal vectors deduced from the dictionary are local optima of the criterion. We illustrate the behavior of the criterion on a 2D example, showing that the local minima correspond to ideal normal vectors when the number of training data is sufficient. We conclude by describing an algorithm that can be used to optimize the criterion in higher dimension.

Index Terms— Sparse representation, dictionary learning, non-convex optimization

1. INTRODUCTION

The efficiency of sparse decompositions in applications highly depends on the match between the dictionary used for the decomposition and the class of processed or analyzed data. Even if appropriate types of dictionaries are known for certain classes of signals, it is often not possible to choose a dictionary a priori, and choice of a good dictionary then requires extensive study of the class of signal under examination. To overcome this difficulty, several methods have been proposed to estimate an appropriate dictionary from a set of training data, in a process commonly referred to as *dictionary learning* (see e.g. [1], [2], [3]). In this paper, we propose a new approach for dictionary learning. In our approach, instead of seeking to model the data directly, the criterion is designed to identify the vectors orthogonal to the subspaces in which the data concentrate.

The authors acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 225913 (project SMALL). MDP is supported by an EPSRC Leadership Fellowship (EP/G007177/1).

After specifying our notations for the dictionary learning problem in Section 1.1, we define the proposed criterion in Section 1.2. We study conditions under which the criterion presents an optimum for ideal normal vectors derived from dictionary in Section 2. We illustrate the behavior of the criterion on some characteristics examples in Section 3, and finally describe the elementary tools that can be used to build an optimization algorithm for the criterion in Section 4.

1.1. Problem setting

We consider a set of training data consisting of N vectors $y_n \in \mathbb{R}^d$, $1 \leq n \leq N$. We suppose that these vectors admit a *sparse* decomposition using an unknown dictionary represented by the $d \times K$ matrix Φ_0, each column k of the matrix being one vector $\phi_k \in \mathbb{R}^d$ of the dictionary. That is to say, each data vector y_n can be written as $y_n = \Phi_0 x_n$, with $x_n \in \mathbb{R}^K$ a sparse coefficient vector, i.e. having few non-zero entries.

These relations can be summarized in convenient matrix notation, by denoting Y the $d \times N$ data matrix whose column n is the vector y_n, and X_0 the $K \times N$ matrix whose column n is the vector x_n. We then have the relation $Y = \Phi_0 X_0$. The problem is then to estimate the dictionary Φ_0 given the data matrix Y.

Just as in blind source separation and independent component analysis, the problem intrinsically suffers from permutation and scaling ambiguities. While the permutation problem is not an issue for us here, to solve the scaling ambiguity, we fix by convention that the columns of Φ_0 must be normed, that is to say that $\|\phi_k\|_2 = 1$ for $1 \leq k \leq K$.

1.2. Criterion definition

With data satisfying $Y = \Phi_0 X_0$ with X_0 sparse, many of the data vectors y_n are a linear combination of a very limited number of the dictionary vectors. This implies that the data will be concentrated on subsets spanned by a limited number of dictionary vectors. In particular for sufficiently sparse data, the vast majority of the data will be contained in the

union of all the hyperplanes (subspaces of dimension $d-1$) spanned by the different possible combinations of $d-1$ dictionary vectors. Therefore, instead of directly searching for the dictionary vectors, we propose to design a criterion intended to identify these subspaces, as a first step towards building a new dictionary learning method. A second step would consist in algebraically recovering the individual dictionary vectors ϕ_k from the collection of hyperplanes they generate.

To identify a subspace in which most of the data are concentrated, we consider a vector $w \in \mathbb{R}^d$ that we assume (without loss of generality) to have unit norm: $\|w\|_2 = 1$. When w is orthogonal to the searched subspace, we have, for most n, $y_n^{\mathrm{T}} w = 0$. Therefore a natural way to identify such a w would be to solve the problem $\min_w \left\|Y^{\mathrm{T}} w\right\|_0$, where $\|x\|_0$ is the number of nonzero entries in the vector x, sometimes referred to as the ℓ^0-norm. But this problem is nonconvex and nonsmooth making it very hard to solve directly. So we replace the preceding ℓ^0-norm by an ℓ^1-norm, defining the criterion C by:

$$C(w) = \left\|Y^{\mathrm{T}} w\right\|_1, \tag{1}$$

and consider the continuous and piecewise smooth problem $\min_w C(w)$. For sufficiently sparse data, we expect that the ideal normal vectors, defined as the vectors normal to the hyperplanes generated by the dictionary vectors, will correspond to local minima of the criterion. More precisely, the ideal normal vectors are vectors w_0 verifying $w_0 \perp \phi_k$ for a given combination of $d-1$ vectors taken from the K vectors of the dictionary. The total number of normal ideal vectors is thus $2\binom{K}{d-1}$.

To study the relevance of the criterion C in equation (1), the following questions will be studied in the next sections:

(a) What is the characterization (necessary and sufficient condition) of local minima of C?
(b) Given training data $Y = \Phi_0 X_0$, are the ideal normal vectors actually local minima of C?
(c) Are there other local minima, which are not associated with ideal normal vectors?
(d) How can one numerically perform the optimization?

2. ANALYSIS OF LOCAL MINIMA

In this section we investigate questions (a) and (b).

2.1. Characterization of local minima

For a given $w \in \mathbb{R}^d$, we define $\Lambda = \{n | y_n^{\mathrm{T}} w \neq 0\}$. We use this set to split the data matrix into two matrices: Y_Λ containing the data vectors nonorthogonal to w and $Y_{\overline{\Lambda}}$ containing the data vectors orthogonal to w. So we define Y_Λ and $Y_{\overline{\Lambda}}$ by $Y_\Lambda = \left[y_n\right], n \in \Lambda$ and $Y_{\overline{\Lambda}} = \left[y_n\right], n \notin \Lambda$.

Theorem 1. *w is a local minimum of C if and only if*

$$\forall w' \perp w, \left|\left(Y_\Lambda \operatorname{sign}\left(Y_\Lambda^{\mathrm{T}} w\right)\right)^{\mathrm{T}} w'\right| < \left\|Y_{\overline{\Lambda}}^{\mathrm{T}} w'\right\|_1.$$

Proof. In the following, the notation $< .,. >_F$ denotes the Frobenius inner product inducing the Frobenius matrix norm.

For two vectors w and w' of $\mathbb{R}^d$ such that $\|w\|_2 = \|w'\|_2 = 1$ and $w^{\mathrm{T}} w' = 0$, we have, if ϵ is sufficiently small, $\operatorname{sign}\left(Y_\Lambda^{\mathrm{T}}(w+\epsilon w')\right) = \operatorname{sign}\left(Y_\Lambda^{\mathrm{T}} w\right)$ and we can write:

$$\begin{aligned}
&C(w+\epsilon w') \\
&= \left\|Y^{\mathrm{T}}(w+\epsilon w')\right\|_1 \\
&= \left\langle Y^{\mathrm{T}}(w+\epsilon w'), \operatorname{sign}\left(Y^{\mathrm{T}}(w+\epsilon w')\right)\right\rangle_F \\
&= \left\langle Y_\Lambda^{\mathrm{T}}(w+\epsilon w'), \operatorname{sign}\left(Y_\Lambda^{\mathrm{T}}(w+\epsilon w')\right)\right\rangle_F \\
&\quad + \left\langle Y_{\overline{\Lambda}}^{\mathrm{T}}(w+\epsilon w'), \operatorname{sign}\left(Y_{\overline{\Lambda}}^{\mathrm{T}}(w+\epsilon w')\right)\right\rangle_F \\
&= \left\langle Y_\Lambda^{\mathrm{T}}(w+\epsilon w'), \operatorname{sign}\left(Y_\Lambda^{\mathrm{T}} w\right)\right\rangle_F \\
&\quad + \left\langle Y_{\overline{\Lambda}}^{\mathrm{T}} \epsilon w', \operatorname{sign}\left(Y_{\overline{\Lambda}}^{\mathrm{T}} \epsilon w'\right)\right\rangle_F \\
&= C(w) + \epsilon \left\langle Y_\Lambda^{\mathrm{T}} w', \operatorname{sign}\left(Y_\Lambda^{\mathrm{T}} w\right)\right\rangle_F + |\epsilon| \left\|Y_{\overline{\Lambda}}^{\mathrm{T}} w'\right\|_1 \\
&= C(w) + \epsilon \left\langle w', Y_\Lambda \operatorname{sign}\left(Y_\Lambda^{\mathrm{T}} w\right)\right\rangle_F + |\epsilon| \left\|Y_{\overline{\Lambda}}^{\mathrm{T}} w'\right\|_1 \qquad (2)
\end{aligned}$$

So w is a local minimum if and only if

$$\forall w' \perp w, \left|\left\langle w', Y_\Lambda \operatorname{sign}\left(Y_\Lambda^{\mathrm{T}} w\right)\right\rangle_F\right| < \left\|Y_{\overline{\Lambda}}^{\mathrm{T}} w'\right\|_1 \quad \square$$

Therefore we have our answer to question (a). Even if this characterization may appear abstract, it will be useful in the following sections. Further work is needed to understand its geometric meaning.

2.2. Study of optimality of ideal normal vectors

To provide a partial answer to question (b), we consider the case when the dictionary vectors ϕ_k form a *basis* of $\mathbb{R}^d$ (i.e. ϕ_k span the whole space and are linearly independent, so $K = d$). Without loss of generality, up to matching row and column permutations of X_0 and Φ_0, we use the following block matrix notation: $\Phi_0 = \begin{bmatrix}\phi_k & \overline{\Phi}_k\end{bmatrix}$ and $X_0 = \begin{bmatrix} x^k & 0 \\ X_k & \overline{X}_k \end{bmatrix}$. The matrix $\overline{\Phi}_k$ is made of the columns ϕ_ℓ, $\ell \neq k$, while x^k is the row vector containing all the non-zero entries of the row k of X_0. The row permutation of X_0 is chosen such that the first row of the new matrix is the row k of the initial matrix, the same permutation being applied to the column of Φ_0, and the column permutation of X_0 is such that all the non-zero entries of the row k appear in x^k on the left of the matrix (cf. [4] or [5] for details of the notation).

We want to identify the conditions in which the criterion C in Equation (1) presents local minima for the ideal normal vectors verifying in the present case a relation of the form $w_0 \perp \phi_\ell, \ell \neq k$.

Lemma 1. *w_0 is a local minimum of C if and only if*

$$\exists d_k, \|d_k\|_\infty < 1, X_k \operatorname{sign}(x^k)^{\mathrm{T}} + \left\|x^k\right\|_1 \overline{\Phi}_k^\dagger \phi_k = \overline{X}_k d_k,$$

where $\overline{\Phi}_k^\dagger$ is the pseudo-inverse of $\overline{\Phi}_k$.

Proof. We note $u(w) = Y_\Lambda \operatorname{sign}\left(Y_\Lambda^{\mathrm{T}} w\right)$. Observe that we have $w' \perp w_0 \Leftrightarrow \exists \beta, w' = \overline{\Phi}_k \beta$. Therefore w_0 is a local minimum if and only if $\forall \beta, \left|\left\langle \overline{\Phi}_k \beta, u(w_0) \right\rangle_F\right| - \left\| Y_{\overline{\Lambda}}^{\mathrm{T}} \overline{\Phi}_k \beta \right\|_1 < 0$ which is equivalent to (cf. [4]):

$$\exists d, \|d_k\|_\infty < 1, \overline{\Phi}_k^{\mathrm{T}} u(w_0) = \overline{\Phi}_k^{\mathrm{T}} Y_{\overline{\Lambda}} d. \tag{3}$$

As $Y = \Phi_0 X_0 = \begin{bmatrix} \phi_k & \overline{\Phi}_k \end{bmatrix} \begin{bmatrix} x^k & 0 \\ X_k & \overline{X}_k \end{bmatrix}$, we have $Y_\Lambda = \phi_k x^k + \overline{\Phi}_k X_k$ and $Y_{\overline{\Lambda}} = \overline{\Phi}_k \overline{X}_k$. Then $Y_\Lambda^{\mathrm{T}} w_0 = \langle \phi_k, w_0 \rangle \, {x^k}^{\mathrm{T}}$ and $\operatorname{sign}\left(Y_\Lambda^{\mathrm{T}} w_0\right) = \pm \operatorname{sign}(x^k)^{\mathrm{T}}$. Thus

$$\begin{aligned} u(w_0) &= \pm \left(\phi_k x^k + \overline{\Phi}_k X_k\right) \operatorname{sign}(x^k)^{\mathrm{T}} \\ &= \pm \left(\left\|x^k\right\|_1 \phi_k + \overline{\Phi}_k X_k \operatorname{sign}(x^k)^{\mathrm{T}}\right) \end{aligned}$$

and $\overline{\Phi}_k^{\mathrm{T}} u(w_0) = \pm \overline{\Phi}_k^{\mathrm{T}} \overline{\Phi}_k \left(X_k \operatorname{sign}(x^k)^{\mathrm{T}} + \left\|x^k\right\|_1 \overline{\Phi}_k^{\dagger} \phi_k\right)$. Replacing expressions of $\overline{\Phi}_k^{\mathrm{T}} u(w_0)$ and $Y_{\overline{\Lambda}}$ in equation (3), we obtain the result. □

Even if it may appear abstract, Lemma 1 can be compared with Theorem 5.1 of [5]. We observe that the two conditions are exactly similar when the dictionary vectors in Φ_0 form an orthonormal basis. When Φ_0 is not an orthonormal basis, the two conditions are different but share a similar form. It would be interesting to study if one of the conditions implies the other. The similarity of the two expressions is of interest, as it is also shown theoretically in [5] that if X_0 is drawn according to a Bernoulli-Gaussian stochastic model, then the condition in Theorem 5.1 of [5] is satisfied with high probability when $N \geq \Gamma \, d \log d$, with Γ a constant. We conjecture that this result extends to condition in Lemma 1, leading to the following conjecture:

Conjecture 1. *If X_0 is drawn according to a Bernoulli-Gaussian stochastic model, then, when $N \geq \Gamma \, d \log d$, with Γ a constant, ideal normal vectors are local minimum of the criterion C with high probability.*

Note that both Lemma 1 and Conjecture 1 are stated when Φ_0 is a basis, *i.e.*, in a setting where standard Independent Component Analysis (ICA, see e.g. [6]) could be applied to learn Φ_0. However, we foresee extensions of the above approach to deal with the case of overcomplete Φ_0. Moreover the proposed analysis gives an order of magnitude of the number of training samples $N \geq \Gamma \, d \log d$ needed to identify Φ_0.

3. EXPERIMENTS

We study numerically the behavior of criterion C in Equation 1 for 2D data in order to get some preliminary indications to answer questions (b) and (c). Motivated by the theoretical results obtained in [5], we draw the coefficients in X_0 according to a Bernoulli-Gauss distribution, that is to say, by denoting $X_0 = (x_{kn})$, that the x_{kn} are independent and identically-distributed random variables with $x_{kn} = \xi_{kn} g_{kn}$, where the ξ_{kn} are indicator variables taking the value 1 with probability p and 0 with probability $1-p$, i.e. $\xi \sim p\, \delta_1 + (1-p)\delta_0$. The variables g_{kn} follow a standard Gaussian distribution, i.e. centered with unit variance. We choose $p = 0.4$.

We fix $\Phi_0 = \begin{bmatrix} \cos\left(\frac{\pi}{2}+1\right) & \cos\left(\frac{\pi}{2}+2\right) & \cos\left(\frac{\pi}{2}+3\right) \\ \sin\left(\frac{\pi}{2}+1\right) & \sin\left(\frac{\pi}{2}+2\right) & \sin\left(\frac{\pi}{2}+3\right) \end{bmatrix}$, which is overcomplete, and consider the value of the criterion C for a vector $w_\alpha = \begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix}$, with $\alpha \in [0, \pi]$. Given Φ_0, the ideal normal vectors are obtained for α equal to 1, 2 and 3. The corresponding results are shown in Figure 1.

For both configurations ($N = 1000$ and $N = 10000$), we observe that the criterion $C\left(w_\alpha\right)$ exhibits clear local minima corresponding to normal ideal vectors. On the one hand, when the number of data vectors is not sufficiently large (case $N = 1000$), we observe that other local minima are found for other values of α. On the other hand, this does not occur when the number of data becomes sufficiently large (case $N = 10000$). These preliminary results reinforce the idea that the criterion can be a valuable tool for subspace identification.

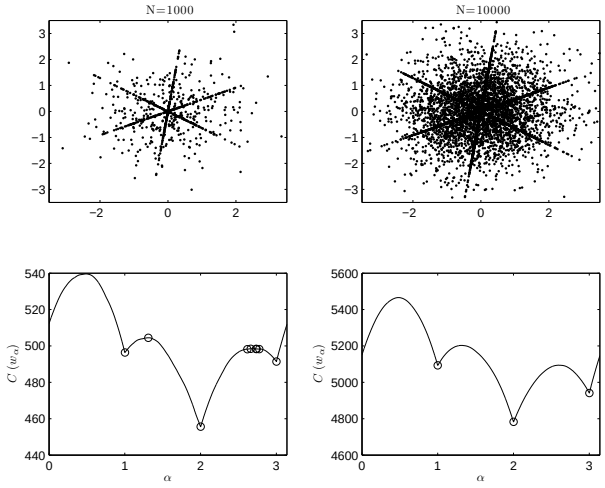

Fig. 1. Examples of synthetic 2D training data and corresponding values of criterion C for a low (N=1000) number of data vectors (left) and for a higher (N=10000) number of data vectors (right). Top: Cloud plot of data y_n. Bottom: plot of the criterion value C (curve) and its local minima (circles).

4. OPTIMIZATION ALGORITHM

We now investigate the numerical optimization of C to answer question (d). We propose an iterative algorithm which alternates between finding the steepest descent direction and using a line search.

Steepest descent: For a fixed $w \in \mathbb{R}^d$, we define the column vector $a = (a_i)$ by $a = Y_\Lambda \operatorname{sign}\left(Y_\Lambda^{\mathrm{T}} w\right)$, the matrix

$B = (b_{ij})$ by $B = Y_\Lambda^{\mathrm{T}}$, and the function f on $\mathbb{R}^d$ by $f(w') = \left|a^{\mathrm{T}} w'\right| - \|Bw'\|_1$.

Considering equation (2), starting from the vector w, the direction of the steepest descent of the criterion will be given by the vector w'_{opt} maximizing the function f. So we want to solve the following optimization problem:

$$\begin{array}{rc} \text{P1:} & \max_{w'} f(w') \\ \text{Subject to} & \|w'\|_2 = 1; w^{\mathrm{T}} w' = 0. \end{array}$$

Two cases should be considered: If $f(w'_{opt}) \leq 0$, then by Theorem 1, w is a local minimum and the descent algorithm must stop. So in this case the value of w'_{opt} is of no interest, and it is only necessary to identify that $f(w'_{opt}) \leq 0$. If $f(w'_{opt}) > 0$ then the value of the criterion can be reduced by following the direction of w'_{opt}, and the knowledge of the exact value of w'_{opt} is needed.

Observation. To find the solution of problem P1, we solve the following quadratic problem with linear constraints:

$$\begin{array}{rc} \text{P2:} & \min_{w',t_j} \|w'\|_2 \\ \text{Subject to} & w^{\mathrm{T}} w' = 0; t_0 - \sum_j t_j = 1 \\ & t_0 \leq \sum_i w'_i a_i; \sum_i w'_i a_i \geq 0 \\ & \forall j \in \{1, \ldots, N - \operatorname{card} \Lambda\}, t_j \geq \sum_i w'_i b_{ij} \\ & \forall j \in \{1, \ldots, N - \operatorname{card} \Lambda\}, t_j \geq -\sum_i w'_i b_{ij}. \end{array}$$

The problem P2 is such that: if P2 as no solution, then $f(w'_{opt}) \leq 0$, indicating that w is optimum; if P2 has a solution for $w' = w''_{opt}$, then $w'_{opt} = \frac{w''_{opt}}{\|w''_{opt}\|_2}$ is solution of P1 and $f(w'_{opt}) > 0$.

Optimization of descent step: We parametrize $w_\theta = \cos\theta\, w + \sin\theta\, w'_{opt}$ and search the minimum $\min_\theta C(w_\theta)$.

Lemma 2. *For $n \in \{1, \ldots, N\}$, let $\rho_n \geq 0$ and θ_n be such that $\langle w, y_n \rangle = -\rho_n \sin\theta_n$ and $\langle w'_{opt}, y_n \rangle = \rho_n \cos\theta_n$, then the minimum of $C(w_\theta)$ is reached for $\theta \in \{\theta_1, \ldots, \theta_N\}$ and $\min_\theta C(w_\theta) = \min_n C(w_{\theta_n})$.*

Proof. We can write $C(w_\theta) = \sum_n \rho_n |\sin(\theta - \theta_n)|$. We define the function $h(\theta) = |\sin\theta|$. For $\theta \neq 0 \bmod \pi$, h is a smooth function twice differentiable and its second derivative satisfies $h''(x) < 0$. So for $x \neq \theta_n \bmod \pi$, the function l defined by $l(\theta) = \sum_n \rho_n |\sin(\theta - \theta_n)|$ is twice differentiable and its second derivative satisfies $l''(x) < 0$, so it cannot have a minimum for $\theta \neq \theta_n \bmod \pi$, and its minimum value is necessarily obtained for a value satisfying $\theta = \theta_n \bmod \pi$. As l is π-periodic, we deduce the result. □

Lemma 2 turns the continuous minimization problem into a finite discrete one.

The complete descent process is summarized in Algorithm 1. Informal experiments indicate that the algorithm performs successfully. We are currently carrying out further experiments to confirm its effectiveness and limitations.

Algorithm 1 Descent algorithm

```
Initialize w
loop
  try to find a solution to problem P2
  if there is a solution w'_opt then
    compute optimal descent step θ_opt (cf. Lemma 2)
    update w: w ← cos θ_opt w + sin θ_opt w'_opt
  else {w is optimum}
    return w
  end if
end loop
```

5. CONCLUSION

We proposed a new criterion for dictionary learning by subspace identification. We characterized its local minima, and empirically demonstrated on a simple example that it enabled ideal normal vectors to be found. We further proposed an algorithm to numerically find local minima of the criterion.

Future work will focus on evaluating the performance of this algorithm for data in higher dimension, and theoretically studying the influence of the number of training data on criterion properties. The use of the proposed approach for greedy dictionary learning using iterative, successive, hierarchic subspace identification will also be investigated.

6. REFERENCES

[1] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[2] K. Kreutz-Delgado, B.D. Murray, J.F.and Rao, K. Engan, T. Lee, and T.J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computations*, vol. 15, no. 2, pp. 349–396, 2003.

[3] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, November 2006.

[4] R. Gribonval and K. Schnass, "Dictionary identifiability from few training samples," in *Proc. 16th EUSIPCO*, Lausanne, Switzerland, August 2008.

[5] R. Gribonval and K. Schnass, "Dictionary identification - sparse matrix-factorisation via ℓ_1-minimisation," preprint (2009), available at http://arxiv.org/abs/0904.4774.

[6] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE. Special issue on blind identification and estimation*, vol. 9, no. 10, pp. 2009–2025, Oct. 1998.